Heuristically Accelerated Reinforcement Learning for Dynamic Secondary Spectrum Sharing

Nils Morozs, *Student Member, IEEE,* Tim Clarke, and David Grace, *Senior Member, IEEE*

***Abstract*—This paper examines how flexible cellular system architectures and efficient spectrum management techniques can be used to play a key role in accommodating the exponentially increasing demand for mobile data capacity in the near future. The efficiency of the use of radio spectrum for wireless communications can be dramatically increased by dynamic secondary spectrum sharing; an intelligent approach that allows unlicensed devices access to those parts of the spectrum that are otherwise underutilised by the incumbent users. In this paper we propose a heuristically accelerated reinforcement learning (HARL) based framework, designed for dynamic secondary spectrum sharing in LTE cellular systems. It utilizes a radio environment map (REM) as external information for guiding the learning process of cognitive cellular systems. System level simulations of a stadium temporary event scenario show that the schemes based on the proposed HARL framework achieve high controllability of spectrum sharing patterns in a fully autonomous way. This results in a significant decrease in primary system quality of service degradation due to the interference from the secondary cognitive systems, compared to a state-of-the-art reinforcement learning solution and a purely heuristic typical LTE solution. The spectrum sharing patterns that emerge by using the proposed schemes also result in remarkable reliability of the cognitive eNodeB on the aerial platform. Furthermore, the novel principle and the general structure of heuristic functions proposed in the context of HARL are applicable to a wide range of self-organization problems beyond the wireless communications domain.**

***Keywords*—Heuristically Accelerated Reinforcement Learning, Spectrum Sharing, Dynamic Spectrum Access**

I. Introduction

One of the fundamental tasks of a cellular system is spectrum management, concerned with dividing the available spectrum into a set of resource blocks or subchannels and assigning them to voice calls and data transmissions in a way which provides a good quality of service (QoS) to the users. Flexible dynamic spectrum access (DSA) techniques play a key role in utilising the given spectrum efficiently in the face of an ever increasing demand for mobile data capacity. This has given rise to novel wireless communication systems such as cognitive radio networks [1] and cognitive cellular systems [2]. Such networks employ intelligent opportunistic DSA techniques that allow them to access licensed spectrum underutilized by the incumbent users.

The classical and most common application of spectrum sharing in cognitive radio networks to date is use of the TV white spaces. Such networks reuse the spectrum allocated to TV broadcasters for other wireless communications, whilst eliminating harmful interference to the incumbent TV receivers, e.g. [3][4]. A more recent problem investigated by researchers, mobile network operators (MNOs) and regulators is LTE and LTE-Advanced spectrum sharing [5]. In many cases LTE spectrum sharing is required by two or more co-primary MNOs. This can be facilitated by an emerging framework known as licensed shared access (LSA) [5]. Here, licenses for the use of LTE spectrum are issued upon agreement for a specific geographical area and required time duration. Another type of LTE spectrum sharing actively investigated within the LTE research community, is resource allocation in heterogeneous networks (HetNets) consisting of LTE femto-cells overlapped by a high power macro-cell, e.g. [6][7]. In these scenarios, the problem is often tackled by using game theory or machine learning principles. The LSA method is a static regulatory approach to spectrum sharing, whereas the HetNet problems normally consider a dynamic scenario, where the same LTE channel is used by both the macro-cell and the femto-cells. Such a problem of dynamic spectrum sharing (DSS) is also investigated in this paper.

An emerging state-of-the-art technique for intelligent DSA and DSS is reinforcement learning (RL); a machine learning technique aimed at building up solutions to decision problems only through trial-and-error [8]. It has been successfully applied to a range of problems and scenarios, such as cognitive radio networks [9], small cell networks [10][11] and cognitive wireless mesh networks [12]. The most widely used RL algorithm in both artificial intelligence and wireless communications domains is Q-learning [13]. Therefore, most of the literature on RL based DSA focuses on Q-learning and its variations, e.g. [11][12][14]. The algorithms developed in this paper are based on distributed Q-learning based DSA. The distributed Q-learning approach has advantages over centralised methods in that no communication overhead is required to achieve the learning objective, and the network operation does not rely on a single computing unit. It also allows for easier insertion and removal of base stations from the network, if necessary. For example, such distributed opportunistic protocols are well suited to temporary event networks and disaster relief scenarios, where rapidly deployable network architectures with unplanned topologies may be required to supplement any existing wireless infrastructure [15].

Although RL algorithms such as Q-learning have been shown to be a powerful approach to problem solving, their common disadvantage is the need for many learning iterations to converge on an acceptable solution. One of the more recent promising solutions to this issue, proposed in the artificial intelligence domain, is the heuristically accelerated reinforcement learning (HARL) approach. Its goal is to speed up RL algorithms, particularly in the multi-agent domain, by guiding the exploration process using additional heuristic

information [16]. In [17], case-based reasoning is used for heuristic acceleration in a multi-agent RL algorithm to assess similarity between states of the environment and to make a guess at what action needs to be taken in a given state, based on the experience obtained in other similar states. In [16], Bianchi et al. prove the convergence of four multi-agent HARL algorithms and show how they outperform the regular RL algorithms. The only example of the HARL approach being applied in the wireless communications domain is the DSA scheme introduced in [11] and used as an integral part of DSS algorithms developed in this paper. There is no evidence in the literature of the HARL approach being applied to a problem of spectrum sharing between two or more separate cellular systems.

The purpose of this paper is to report on the novel application of HARL to the problem of dynamic secondary spectrum sharing. The proposed framework uses a dynamic spectrum database, known as the radio environment map (REM), as heuristic acceleration to mitigate poor temporal performance of RL algorithms applied to DSS problems. Furthermore, the principles and features of the proposed technique aim to be generally applicable to a wide range of learning problems beyond the wireless communications domain. In previous work on combining RL and dynamic spectrum databases, e.g. REMs, researchers have considered employing RL algorithms solely for obtaining information that can be stored in these databases, e.g. [18][19]. There is no evidence of previous work in the literature on using REM databases to enhance the performance of RL based DSA and DSS algorithms.

The rest of the paper is organised as follows: Section II introduces the spectrum sharing problem investigated in this paper. Section III explains the principles behind RL and HARL based DSA. In Section IV we propose a novel HARL framework and show how it can be applied to the DSS problem in hand. Section V evaluates the performance of the proposed schemes by simulating a large scale LTE spectrum sharing scenario. The conclusions are given in Section VI.

II. The Spectrum Sharing Problem

One of the scenarios currently considered in the EU FP7 ABSOLUTE project is a temporary cognitive cellular infrastructure that is deployed in and around a stadium to provide extra capacity and coverage to the mobile subscribers and event organizers involved in a temporary event, e.g. a football match or a concert [20]. This scenario is depicted in Fig. 1, where a small cell network is deployed inside the stadium to provide ultra high capacity density to the event attendees, and an eNodeB (eNB) on an aerial platform is deployed above the stadium to provide wide area coverage. Previous work on this scenario has only considered spectrum sharing between the stadium network and the primary eNBs (PeNBs) using a distributed Q-learning algorithm explained in Subsection III-B [21]. Whereas, the problem investigated in this paper considers dynamic spectrum sharing between the small cell network, the PeNBs, and the aerial eNB (AeNB).This is a more complex problem that motivates the development of novel RL based self-organisation algorithms presented in Section IV.

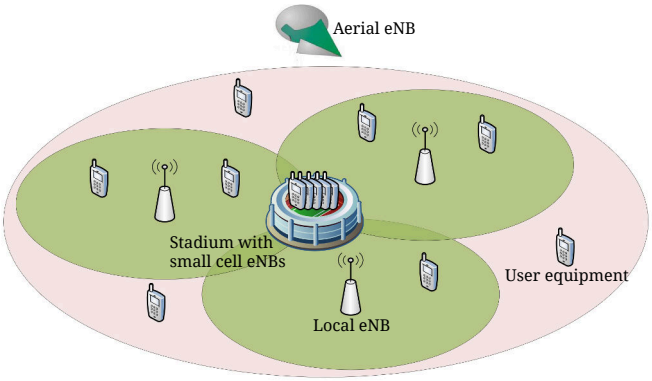


Fig. 1. Stadium temporary event scenario

The cognitive small cells and the AeNB have secondary access to a 20 MHz LTE channel, also used by a network of 3 local PeNBs. The goal of the small cell network and the AeNB is to use distributed machine intelligence methods to form a self-organizing heterogeneous cellular system which reuses the LTE spectrum of the local primary LTE network.

III. Cognitive Dynamic Spectrum Access

In order to discuss secondary spectrum sharing, the DSA mechanism for scheduling resources of the cognitive cellular system alone needs to be introduced first. This section presents the concepts of RL, stateless Q-learning, heuristically accelerated RL (HARL), and explains the details of the HARL based cognitive DSA algorithm designed for the secondary system, initially without considering the presence of a primary system.

A. Reinforcement Learning

RL is a model-free type of machine learning which is aimed at learning the desirability of taking any available action in any state of the environment only through trial-and error [8]. This desirability of an action is represented by a numerical value known as the Q-value - the expected cumulative reward for taking a particular action in a particular state, as shown in the equation below:

$$Q(s,a) = E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \quad (1)$$

where $Q(s,a)$ is the Q-value of action a in state s, r_t is the numerical reward received t time steps after action a is taken in state s, T is the total number of time steps until the end of the learning process or episode, and $\gamma \in (0,1)$ is a discount factor.

The job of an RL algorithm is to estimate $Q(s,a)$ values for every action in every state, which are then stored in an array known as the Q-table. In some cases where an environment does not have to be represented by states, only the action space and a 1-dimensional Q-table $Q(a)$ can be considered [22]. The job of an RL algorithm then becomes simpler, it

aims to estimate an expected value of a single reward for each action available to the learning agent:

$$Q(a) = E[r_t] \tag{2}$$

B. Stateless Q-learning

One of the most successful and widely used RL algorithms is Q-learning. In particular, a simple stateless variant of this algorithm, as formulated in [22], has been shown to be effective for several distributed DSA learning problems, e.g. [21][23].

Each eNB maintains a Q-table $Q(a)$ such that every subchannel a has an expected reward or Q-value associated with it. The Q-value represents the desirability of assigning a particular subchannel to a file transmission. Upon each file arrival, the eNB either assigns a subchannel to its transmission or blocks it if all subchannels are occupied. It decides which subchannel to assign based on the current Q-table and the greedy action selection strategy described by the following equation:

$$\hat{a} = \underset{a}{\operatorname{argmax}}(Q(a)),\ a \in A', A' \subset A \tag{3}$$

where $\hat{a}$ is the subchannel chosen for assignment out of the set of currently unoccupied subchannels A', $Q(a)$ is the Q-value of subchannel a, and A is the full set of subchannels.

The values in the Q-tables are initialised to zero, so all eNBs start learning with equal choice among all available subchannels. A Q-table is updated by the corresponding eNB each time it attempts to assign a subchannel to a file transmission in the form of a positive or a negative reinforcement. The recursive update equation for stateless Q-learning, as defined in [22], is given below:

$$Q(a) \leftarrow (1-\alpha)Q(a) + \alpha r \tag{4}$$

where $Q(a)$ represents the Q-value of the subchannel a, r is the reward associated with the most recent trial and is determined by a reward function, and $\alpha \in [0, 1]$ is the learning rate parameter which weights recent experience with respect to previous estimates of the Q-values.

The reward function, which is generally applicable to a wide range of RL problems and which has been successfully applied to DSA problems in the past [9][24], returns two values:

- $r = -1$ (negative reinforcement), if the file transmission failed due to an insufficient Signal-to-Interference-plus-Noise Ratio (SINR) on the selected subchannel.
- $r = 1$ (positive reinforcement), if the file is successfully transmitted, i.e. SINR did not drop below the transmission threshold.

The choice of the learning rate value for this type of distributed Q-learning based DSA problems is thoroughly investigated in [24]. The best performance is achieved by using the Win-or-Learn-Fast (WoLF) variable learning rate principle [25] described by (5), where a lower value of α is used for successful trials (when $r = 1$), and a higher value of α is used for failed trials ($r = -1$). In this way, the learning agents are learning faster when "losing" and more slowly when "winning".

$$\alpha = \begin{cases} 0.01 & r = 1 \\ 0.1 & r = -1 \end{cases} \tag{5}$$

C. Heuristically Accelerated Reinforcement Learning

A common disadvantage of machine learning algorithms, such as distributed Q-learning described in the previous subsection, is that they are normally used to learn solutions only through trial-and-error with no prior knowledge of the problem in hand. Consequently, it takes a large number of trials for them to learn acceptable solutions. This is undesirable in real-time applications such as DSA in cellular systems. An emerging technique to mitigate this poor initial performance problem is the HARL approach, where additional heuristic information is used to guide the exploration process [16].

Fig. 2 shows our block diagram representation of the processes involved in HARL. It demonstrates that HARL is an extension of regular RL algorithms. The unfilled blocks and solid lines constitute a block diagram of regular RL, whereas dashed lines and shaded blocks indicate the additional functionality afforded by the heuristic acceleration.

The role of the inner RL loop is to learn a good policy to be used by the learning agent. It achieves this goal by observing the actions taken by the learning agent, sampling the outputs caused by them, and directly estimating (updating) the entries in the Q-table. The role of the policy is to map every state of the environment to the most appropriate action that can be taken in that state. It can be derived from the estimated Q-table and used for decision making. In the context of the DSA problem, the output of interest is whether or not a file transmission is blocked or interrupted, and the action is the piece of resources allocated it.

The key additional element provided by HARL is the derivation of a heuristic policy. According to [16], a heuristic policy is derived from additional knowledge, either external or internal, which is not included in the learning process. Generally, the goal of the heuristic policy $H_t(s, a)$ is to influence the action choices of a learning agent, i.e. to modify its current policy $\pi_t(s)$ in a way which would accelerate the learning process. The format and dimensions of $H_t(s, a)$ should be compliant with the Q-table used by the given learning agent, such that its new combined policy $\pi_t^c(s)$ can be derived using the following equation:

$$\pi_t^c(s) = \underset{a}{\operatorname{argmax}}(Q_t(s, a) + H_t(s, a)) \tag{6}$$

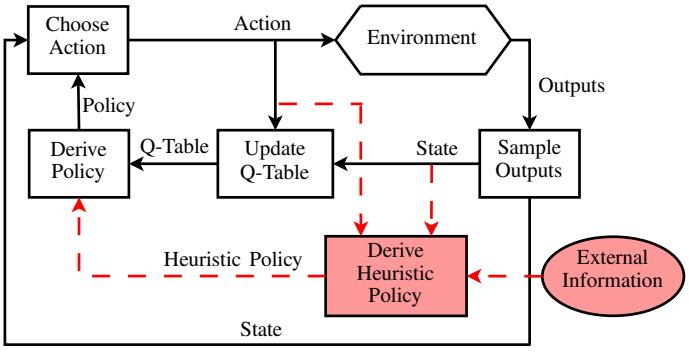


Fig. 2. Block diagram of heuristically accelerated reinforcement learning

where $\pi_t^c(s)$ is the combined policy of the given learning agent for state s at time t based on its Q-table $Q_t(s,a)$ and the heuristic policy $H_t(s,a)$. If $H_t(s,a)$ is always zero, the algorithm becomes a regular RL algorithm.

D. Heuristic Acceleration Using ICIC

The only existing HARL based DSA scheme is known as distributed ICIC accelerated Q-learning (DIAQ), proposed in [11]. It uses inter-cell interference coordination (ICIC) signalling in the LTE downlink as heuristic acceleration for a distributed stateless Q-learning algorithm described in Subsection III-B. It achieves dramatic improvements in initial and steady-state QoS, as well as in learning convergence rate, in a cognitive cellular system with dedicated spectrum.

The format of the messages exchanged between eNBs using ICIC in the LTE downlink is standardized by the 3GPP and referred to as the Relative Narrowband Transmit Power (RNTP) indicator [26]. It contains a bitmap which indicates on which resource blocks an eNB is planning to transmit at high power by setting their corresponding bits to *1*, i.e. on which resource blocks it is likely to cause interference in adjacent cells. For example, in a case where a 20 MHz LTE channel has 25 subchannels, the length of an RNTP message is 100 bits or 25 hexadecimal characters [26]. Since every subchannel consists of 4 adjacent resource blocks, every group of 4 bits (i.e. every hexadecimal character) in an RNTP message describes a particular subchannel. For example, if an eNB is planning to use high transmit power on a given subchannel, its corresponding bits in the RNTP message are *1111* or *0xF*.

The choice of the RNTP threshold used to decide whether a given transmit power is high or low is set to -3 dB with respect to the average transmit power in a cell [11]. To avoid excessive signalling requirements, the time interval between the ICIC message exchanges is assumed to be 20 ms [27].

When a request for a new file transmission is received, the eNB starts by aggregating the latest RNTP messages from its neighbours into an ICIC bitmask using a bitwise *OR* operation, as described by the following equation:

$$Mask_{ICIC} = \bigcup_{n=1}^{N} RNTP_n \tag{7}$$

where $Mask_{ICIC}$ is a 25 hexadecimal character string representing the subchannels reserved by *any* of the neighbouring eNBs by *0xF*, and representing the "safe-to-use" subchannels by *0x0*, $RNTP_n$ is a 25 hexadecimal character RNTP message of the n'th neighbouring eNB, and N is the total number of neighbouring eNBs.

After creating the ICIC mask, the eNB creates a heuristic policy $H_{ICIC}(a)$ using the following principle:

$$H_{ICIC}(a) = \begin{cases} h_{ICIC} & Mask_{ICIC}(a) = 0xF \\ 0 & Mask_{ICIC}(a) = 0x0 \end{cases} \tag{8}$$

where $H_{ICIC}(a)$ is the heuristic policy value of subchannel a, $h_{ICIC} < q_{min} - q_{max}$ is a fixed negative number with greater amplitude than the difference between the minimum (q_{min}) and the maximum (q_{max}) possible values in the Q-tables, and $Mask_{ICIC}(a)$ is a character in the ICIC mask that corresponds to subchannel a. $H_{ICIC}(a)$ can then be employed to create a temporary masked Q-table $Q_m(a)$ using (9), which in turn is used for heuristically guided decision making, whilst a normal learning process is taking place using the original Q-table $Q(a)$.

$$Q_m(a) = Q(a) + H_{ICIC}(a) \tag{9}$$

By using such a heuristic policy $H_{ICIC}(a)$, the eNB is guaranteed to prioritise the subchannels marked as "safe" by the ICIC bitmask before the "unsafe" subchannels by shifting the Q-values of the latter to the bottom of the Q-table, whilst still preserving their respective order in terms of the Q-values (due to the fixed value of h_{ICIC}).

IV. HARL for Dynamic Spectrum Sharing

The stadium temporary event spectrum sharing scenario described in Fig. 1 consists of a network of primary eNBs (PeNBs) operating in a suburban area and a secondary cognitive cellular system that itself consists of two separately operating entities - an aerial eNB (AeNB) for wide area coverage and a small cell network for high capacity density inside the stadium.

A study in [21] has demonstrated that successful dynamic spectrum sharing between a low power stadium small cell system and a relatively high power local PeNB infrastructure can be facilitated using an independent distributed Q-learning algorithm implemented in the former. This is largely because the interference between the two systems is attenuated by the stadium shell. However, the scenario investigated in this paper also involves an AeNB serving line-of-sight (LoS) users both inside and outside the stadium. Therefore, it presents two additional challenges - spectrum sharing between the PeNBs and the AeNB, and spectrum sharing between the AeNB and the stadium small cell network.

Our proposed way of achieving these two spectrum sharing tasks is to use a small scale database, referred to as the radio environment map (REM) [28], to continuously monitor and store the information about spectrum usage of the PeNBs and the AeNB. In this way, the AeNB has a means to avoid interfering with the primary system, and the small cell network can avoid interfering with the AeNB. This type of setup is depicted in Fig. 3, which is a classical way of achieving

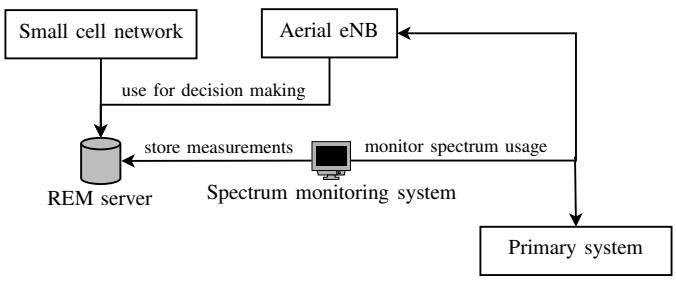


Fig. 3. Secondary spectrum sharing using a spectrum monitoring system and a radio environment map (REM)

coexistence between cognitive radio networks and primary spectrum users, especially in the TV white space context [3][4].

The task of the spectrum monitoring system with a REM database is to detect the occupancy of the spectrum resources used by the PeNBs and the AeNB. It is then possible to estimate the probability of spectrum occupancy at every eNB on every individual subchannel that, in turn, can be used to influence the spectrum assignment decisions of the secondary systems.

A. Spectrum Monitoring

One way of implementing reliable spectrum monitoring in such LTE cellular systems is for the primary system to grant the secondary system access to its ICIC signals. This approach was introduced in [21] to investigate the need for spectrum awareness in a stadium cognitive small cell network. The ICIC signals standardized for the LTE downlink are explained in Subsection III-D. In this way, the binary spectrum occupancy information about the PeNBs and the AeNB could be logged at the REM server and used to make predictions about the spectrum availability. Such a protocol is easily implementable, especially if both systems are controlled by the same mobile network operator (MNO). However, if the secondary cognitive network is not controlled by the primary system's MNO, it may not be allowed to access the ICIC signals of the primary system. In such cases, dynamic spectrum monitoring could be achieved by deploying a sensor network around the stadium to detect spectrum usage of every PeNB and AeNB, e.g. using an algorithm for multiple signal classification [29].

Regardless of the detection mechanism, the algorithms proposed in this section assume that the spectrum monitoring system is able to periodically detect whether or not a particular subchannel is being used by a particular PeNB or AeNB. It is designed to return 1 if it is currently occupied, or 0 otherwise.

B. Spectrum Occupancy Estimation

Given this mechanism for obtaining a stream of binary spectrum occupancy data, it is then important to estimate the probability of subchannel occupancy at every observed eNB, i.e. a probability of a particular subchannel being occupied at a particular eNB based on the previous observations.

A simple and appropriate way of tracking the mean of a data sequence, whilst simultaneously giving more recent observations higher weight compared to older estimates, is the exponentially weighted moving average (EWMA) method, e.g. [30]. It can be calculated using the following recursive equation:

$$y \leftarrow (1-\lambda)y + \lambda x \tag{10}$$

where y is the mean estimate of the data sequence x, and λ is a factor which controls how quickly the estimated mean adapts to new observations. The role of λ in EWMA estimation is identical to that of the learning rate α in the Q-learning update formula from (4). In fact, comparing Equations (4) and (10) demonstrates that stateless Q-learning is, in fact, an EWMA estimation algorithm of the rewards received by a learning agent.

We propose adapting the EWMA method to estimate the probability of subchannel occupancy $p(occupied)$ in the following way:

$$p(occupied) \leftarrow (1-\lambda)p(occupied) + \lambda b,\ b \in \{0,1\} \tag{11}$$

where b is a current binary subchannel occupancy measurement, i.e. $b = 1$ if the given subchannel is occupied, $b = 0$ if it is not. In this way, the EWMA equation is used to estimate the mean of a stream of 1's and 0's, representing $p(occupied) \in [0,1]$.

C. REM Based Heuristic Function

A threshold P_{min} to determine whether a particular subchannel should be avoided, based on an estimate of $p(occupied)$, can then be defined to obtain the following heuristic function:

$$H_{REM}(a) = \begin{cases} h_{REM} & p_a(occupied) \geq P_{min} \\ 0 & p_a(occupied) < P_{min} \end{cases} \tag{12}$$

where $H_{REM}(a)$ is the value of the REM based heuristic function for subchannel a, $p_a(occupied)$ is the EWMA estimate of $p(occupied)$ for subchannel a, h_{REM} is a fixed negative value which shifts the Q-values of the undesirable subchannels down, such that the others are prioritized before them. This heuristic function follows the same principle of shifting Q-values as the one used in DIAQ (see Subsection III-D).

Such a heuristic function $H_{REM}(a)$ aims to guide the learning process of the cognitive eNBs in a direction desirable for secondary spectrum sharing. The small cell eNBs can coexist with the AeNB by applying the heuristic function from (12) to the AeNB subchannel occupancy observations, hereafter referred to as $H_{REM}^{AeNB}(a)$. Whereas the AeNB can coexist with the PeNBs by applying the same principle to PeNB subchannel occupancy observations. In this case, since the wide area coverage AeNB is going to interfere with all PeNBs in the area of interest, the probability of subchannel a being occupied by any PeNB is obtained by calculating the sum of $p_a(occupied)$ values of every individual PeNB:

$$p_a^{any\ PeNB}(occupied) = \sum_{n=1}^{N} p_a^{n^{th}\ PeNB}(occupied) \tag{13}$$

where N is the total number of PeNBs. The REM based heuristic function from (12) can then be calculated using $p_a^{any\ PeNB}(occupied)$, hereafter referred to as $H_{REM}^{PeNBs}(a)$.

D. Superimposed Heuristic Functions

With the introduction of the REM based heuristic function for secondary spectrum sharing, a framework for using several heuristic functions simultaneously is required. For example, in addition to using an ICIC based heuristic function $H_{ICIC}(a)$ introduced in Subsection III-D for internal dynamic spectrum access, the small cell eNBs are now also required to share spectrum with the AeNB using another heuristic function

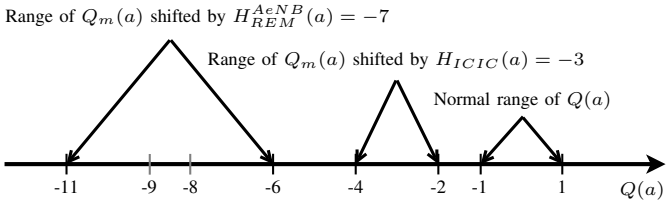


Fig. 4. The effect of superimposed heuristic functions $H_{ICIC}(a) \in \{0, -3\}$ and $H_{REM}^{AeNB}(a) \in \{0, -7\}$ on the range of masked Q-table values

$H_{REM}^{AeNB}(a)$, such that their masked Q-tables $Q_m(a)$ could be constructed using the following principle:

$$Q_m(a) = Q(a) + H_{ICIC}(a) + H_{REM}^{AeNB}(a) \tag{14}$$

where $Q(a) \in [-1, 1]$ is an original Q-table of a given eNB maintained using the stateless Q-learning algorithm described in Subsection III-B. There, two heuristic functions $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ have to be superimposed to modify a learning eNB's policy, such that it incorporates both ICIC and REM information into its learning process.

We propose a method where every new heuristic function superimposed on the Q-table splits the Q-values into two non-overlapping regions, as shown in Fig. 4. The normal range of Q-values $Q(a)$ maintained by the stateless Q-learning algorithm from Subsection III-B is $[-1, 1]$. If the h_{ICIC} parameter of the $H_{ICIC}(a)$ heuristic function is -3, it shifts $Q_m(a)$ values of disapproved subchannels into a non-overlapping region of $(Q(a) - 3) \in [-4, -2]$, thus prioritizing them below the subchannels with $Q_m(a) \in [-1, 1]$. If another heuristic function $H_{REM}^{AeNB}(a)$ is used and its h_{REM} constant is -7, it will split $Q_m(a)$ into two regions - $Q_m(a) \in [-4, 1]$ and $(Q_m(a) - 7) \in [-11, -6]$. In this way, the subchannels disapproved by $H_{REM}^{AeNB}(a)$ are guaranteed to be prioritized below any other subchannel. This approach allows an unlimited number of further heuristic functions superimposed on top of each other, as long as their respective importance is known. For example, in this case we prioritize $H_{REM}^{AeNB}(a)$ responsible for spectrum sharing above $H_{ICIC}(a)$ responsible for internal stadium network DSA by setting $h_{REM} < h_{ICIC}$.

E. Q-Value Based Admission Control

The HARL algorithm required for the AeNB to coexist with the primary system only includes one heuristic function $H_{REM}^{PeNBs}(a)$, since it is a separately controlled entity with no ICIC-compatible neighbouring base stations. Therefore, it uses the following masked Q-table for guiding its learning process:

$$Q_m(a) = Q(a) + H_{REM}^{PeNBs}(a) \tag{15}$$

However, another important aspect of secondary spectrum sharing is the primary user protection [31], i.e. making sure the secondary system, in this case the AeNB, does not produce harmful interference for the primary system, in our case the users connected to the PeNBs. A technique that could be easily and effectively embedded into the HARL framework developed in this paper, i.e. where $H_{REM}^{PeNBs}(a)$ shifts part of the Q-values by a fixed negative number h_{REM}^{PeNBs}, is Q-value based admission control (Q-AC) introduced in [23]. There, a Q-value threshold q_{AC} is defined, such that:

$$A_{allowed} = \{a \mid a \in A' \ \wedge \ Q(a) \geq q_{AC}\} \tag{16}$$

where A' is the set of currently unoccupied subchannels, i.e. those available for assignment, and $A_{allowed} \subset A'$ is the set of subchannels allowed for assignment based on the admission threshold q_{AC}. In this way, the subchannels with $Q(a) < q_{AC}$ are never assigned to data transmissions, which are blocked instead.

The value of q_{AC} can be chosen such that:

$$q_{max} - h_{REM}^{PeNBs} < q_{AC} < q_{min} \tag{17}$$

where q_{min} and q_{max} are the minimum and the maximum possible value of $Q(a)$ respectively. In this way, the subchannels disapproved by the heuristic function $H_{REM}^{PeNBs}(a)$ will be forbidden to be assigned at the AeNB, due to their Q-values being shifted below q_{AC}, thus guaranteeing protection of the PeNBs from secondary interference.

F. HARL Algorithms for Spectrum Sharing

Algorithms 1 and 2 summarize the HARL schemes for dynamic secondary spectrum sharing developed in this section. Algorithm 1 shows the sequence of steps in the distributed REM and ICIC accelerated Q-learning (DRIAQ) scheme, designed for stadium small cells to mitigate interference among themselves and the AeNB, using two superimposed heuristic functions. Algorithm 2 shows the REM accelerated Q-learning algorithm with Q-value based admission control (RAQ-AC), designed for the AeNB to share spectrum and avoid interference with the primary system.

Lines {2, 8, 9} of Algorithm 1 and lines {2, 8-12, 14} of Algorithm 2 are specific to the novel HARL schemes developed in this section. If they are removed and $Q_m(a)$ is substituted by $Q(a)$, the algorithms are simplified down to stateless Q-learning from Subsection III-B.

Algorithm 1 Distributed REM and ICIC accelerated Q-learning (DRIAQ) for stadium small cells

1: Initialise Q-table to all zeros
2: Set $h_{ICIC} = -3$ and $h_{REM}^{AeNB} = -7$
3: **while** eNB is on **do**
4: Wait for a file arrival
5: **if** all subchannels are occupied **then**
6: Block transmission
7: **else**
8: Update $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ based on latest ICIC and REM information, using (8) and (12)
9: Combine $Q(a)$ with $H_{ICIC}(a)$ and $H_{REM}^{AeNB}(a)$ into a masked Q-table $Q_m(a)$ using (14)
10: Assign the best subchannel using $Q_m(a)$ and (3)
11: Observe the outcome, calculate the reward $r = \pm 1$
12: Update $Q(a)$ using (4)
13: **end if**
14: **end while**

Algorithm 2 REM accelerated Q-learning with Q-value based admission control (RAQ-AC) for the aerial eNB

Initialise Q-table to all zeros
Set $h_{REM}^{PeNBs} = -7$ and $q_{AC} \in (-6, -1)$ as shown in (17)
while eNB is on **do**
  Wait for a file arrival
  if all subchannels are occupied **then**
    Block transmission
  else
    Update $H_{REM}^{PeNBs}(a)$ based on latest REM information, using (12)
    Combine $Q(a)$ with $H_{REM}^{PeNBs}(a)$ into a masked Q-table $Q_m(a)$ using (15)
    if all subchannels with $Q_m(a) \geq q_{AC}$ are occupied **then**
      Block transmission
    else
      Assign the best subchannel using $Q_m(a)$ and (3)
    end if
    Observe the outcome, calculate the reward $r = \pm 1$
    Update $Q(a)$ using (4)
  end if
end while

G. Choice of Parameters

The final details required to complete the design of the REM and the REM based heuristic functions are the values of the EWMA algorithm parameter λ from (10) and the probability of subchannel occupancy threshold P_{min} for $H_{REM}^{AeNB}(a)$ and $H_{REM}^{PeNBs}(a)$ as used in (12). We propose using $P_{min} = \lambda$ and $\lambda = 0.008$, while the REM is updated every 200 ms, which is frequent enough to capture the traffic variations of the PeNBs and the AeNB, yet not too frequent to introduce a large overhead of additional REM information that has to be broadcast to all cognitive eNBs. However, other values can be used for these parameters without the loss of generality.

The value of λ is chosen based on the rate of decay of a $p_a(occupied)$ estimate, e.g. the time it would take for a once heavily used subchannel to be assumed unused, if the eNB of interest stopped using it. For example, if $p_a(subchannel) = 0.99$ and afterwards subchannel a is not used for 600 consecutive REM updates, i.e. 2 minutes, the new $p_a(occupied)$ estimate, based on (11), is the following:

$$p_a(occupied) = 0.99 \times (1 - \lambda)^{600} = 0.00799 \qquad (18)$$

which is just below $P_{min} = \lambda = 0.008$. Therefore subchannel a would no longer be undesirable for secondary reuse, based on the heuristic function from (12). This value of λ is high enough to be applicable in dynamic environments where the monitored spectrum usage patterns change over time, yet not high enough to dismiss valuable historical spectrum usage information too quickly. This trade-off between the speed and accuracy of the EWMA algorithm, controlled by the λ parameter, is essential and must be carefully considered, e.g. using numerical examples such as the one described in (18).

The value $P_{min} = \lambda$ is proposed because it is crucial that, if interference is detected on a previously unused subchannel with $p(occupied) = 0$, the new estimate of $p(occupied)$ is such that this subchannel is recognised as busy straightaway. In this case the $p(occupied)$ estimate will change from 0 to $\lambda = P_{min}$ which is high enough to be flagged by the REM based heuristic function described by (12).

V. Simulation Results

The spectrum sharing problem described in Section II involves an AeNB and a network of small cell eNBs that have to share spectrum among themselves and with a primary system of local eNBs operating in the area.

The primary system is assumed to employ a dynamic ICIC scheme, where all three PeNBs exchange their current spectrum usage as RNTP messages every 20 ms, and exclude the subchannels currently used by the other two PeNBs from their available subchannel list [11][27]. We assume that they always try to assign an available subchannel with the lowest index if any, e.g. they always scan the availability of the subchannels in the same order from the 1st subchannel to the last. In this way, the primary network would make its spectrum usage less random and more appropriate for the cognitive cellular system to share, which is in the interests of both the primary and the secondary system. However, the dynamic spectrum sharing schemes developed for the secondary systems in Section IV do not assume this and would also work regardless of the spectrum management strategy of the primary system.

The results of implementing the following three schemes in the secondary cognitive system are discussed in this section:

- "Dynamic ICIC" - all systems use ICIC signalling as described above for the primary system. The stadium eNBs receive ICIC messages from the AeNB and from their neighbouring small cells. They only report subchannels used at a Tx power above -3 dB with respect to the average power in the cell, and choose randomly among the subchannels deemed "safe". The AeNB randomly assigns subchannels not used by the primary system, based on the ICIC messages of the latter.
- "DIAQ + Q-learning" - all networks are working independently. The stadium network employs the DIAQ scheme introduced in Subsection III-D, and the AeNB is using stateless Q-learning from Subsection III-B. This scheme represents a state-of-the-art RL solution to the spectrum sharing problem.
- "DRIAQ + RAQ-AC" - the combination of novel HARL based schemes developed in Section IV and summarized in Algorithms 1 and 2.

A. Stadium Temporary Event Network

The stadium small cell network architecture is depicted in Fig. 5, where the users are located in a circular spectator area 53.7 - 113.7 m from the centre of the stadium. The spectator area is covered by 78 eNBs arranged in three rings at 1 m height, e.g. with antennas attached to the backs of the seats or to the railings between the different row levels. Seat width is assumed to be 0.5 m, and the space between rows - 1.5 m, which yields the total capacity of 43,103 seats.

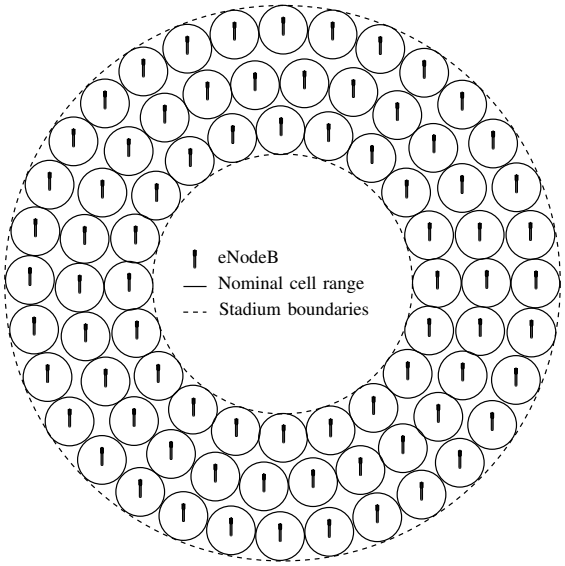


Fig. 5. Stadium network architecture

The cognitive small cell network and the AeNB, located above the stadium centre point at 300 m altitude, have secondary access to a 20 MHz LTE channel also used by the primary network. It consists of 3 PeNBs whose coordinates, with respect to the centre point of the stadium, are $(-600, -750)$, $(100, 750)$ and $(750, -800)$ m. Therefore, the goal of the cognitive small cell and aerial eNBs is to efficiently utilize the 20 MHz LTE channel, normally reserved for the PeNBs, whilst avoiding interference with them.

500 user equipments (UEs) are randomly distributed outside the stadium, in the circular area from the stadium boundary (5 m from the radius of the last row) to 1.5 km away from the stadium centre point. 25% of the stadium capacity is filled with randomly distributed wireless subscribers, i.e. $\approx$ 10,776 UEs. The offered traffic is 20 Mb/s outside of the stadium and 1 Gb/s inside. All simulations last 2,000,000 transmissions, most of which take place inside the densely populated stadium. This corresponds to $\approx$2 hours. The parameters and assumptions of the simulation model are listed in Table I.

B. Spectrum Occupancy Analysis

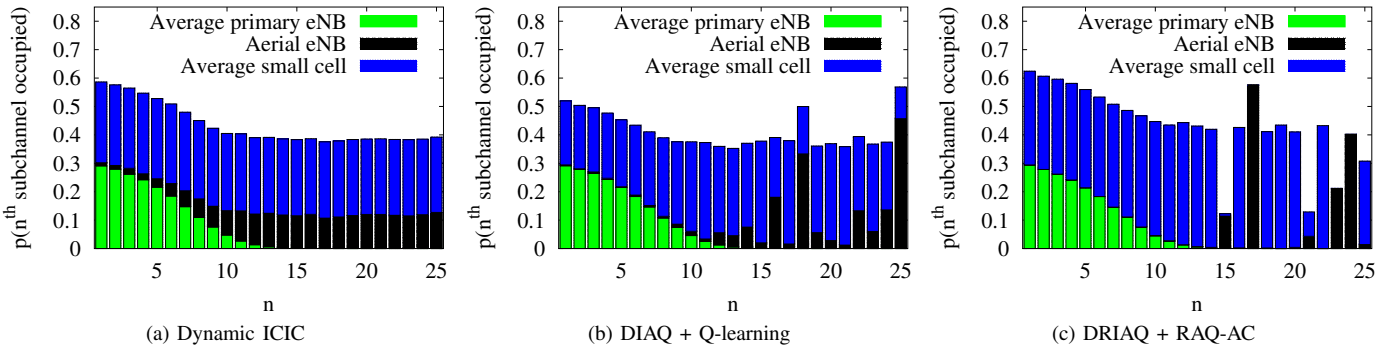
Fig. 6 shows the subchannel occupancy distributions of the PeNBs, the AeNB, and the small cell eNBs using three different spectrum sharing strategies described at the beginning of this section. These distributions were calculated by measuring the amount of time every eNB spent occupying every subchannel and dividing it by the total simulation time.

Fig. 6a shows that in the case of "dynamic ICIC" implemented in all systems, the reverse relationship between the spectrum mostly used by the AeNB and that preferred by the primary system is observed, demonstrating the effect of frequent ICIC signalling between the two. It also shows that the small cell network uses the whole spectrum approximately uniformly. Fig. 6b demonstrates the difference made by introducing distributed Q-learning into the DSS process. The two challenging spectrum sharing relationships associated with this scenario tend to be addressed through distributed machine intelligence. The AeNB learns to avoid using the primary spectrum more than the "dynamic ICIC" approach, whilst the small cell eNBs tend to learn to use the subchannels preferred by the AeNB less than the others, i.e. they learn to avoid interfering with the AeNB, since it often results in blocked and interrupted file transmissions.

Fig. 6c shows how the novel heuristically accelerated approach further improves the autonomously emerging spectrum sharing pattern by strictly guiding the learning process of the AeNB to avoid interfering with the PeNBs, and discouraging

TABLE I. NETWORK MODEL PARAMETERS AND ASSUMPTIONS

Parameter	**Value**
Channel bandwidth	20 MHz: 100 LTE virtual resource blocks (VRBs)
Subchannel bandwidth	4 VRBs: 4 $\times$ 180 kHz [26]
Frequency band	2.6 GHz
UE receiver noise floor	94 dBm (290 K temperature, 20 MHz bandwidth, 7 dB noise figure)
Stadium propagation	WINNER II B3 [32]
Outdoor propagation	WINNER II C1 [32]
Propagation between stadium and outdoors	Combined WINNER II C4 with C1 term [32]
Propagation between AeNB and ground	Free space + 8dB log-normal shadowing
Traffic model	3GPP FTP Traffic Model 1 [33], file size - 4.2 Mb
Retransmissions	Uniform random back-off between 0 and 960 ms [34]
Link model	3GPP Truncated Shannon Bound model [35]
Primary eNB Tx power	10 dBW
Assumptions	
UEs inside the stadium are associated with a small cell or aerial eNB with a minimum estimated downlink pathloss, based on the Reference Signal Received Power (RSRP)	
UEs outside the stadium are associated with a primary or aerial eNB based on the strongest RSRP. The reference signal Tx power of the primary eNB is 13 dB higher than that of the AeNB	
Cognitive small cell and aerial eNBs employ open loop power control, using a constant Rx power of -74 dBm (20 dB Signal-to-Noise Ratio)	
The minimum Signal-to-Interference-plus-Noise Ratio (SINR) allowed to support data transmission is 1.8 dB	
One subchannel (4 VRBs) is allocated to every data transmission	



(a) Dynamic ICIC (b) DIAQ + Q-learning (c) DRIAQ + RAQ-AC

Fig. 6. Subchannel occupancy of primary eNBs, aerial eNB and small cells using different spectrum sharing schemes

the small cell eNBs from exploring and assigning the subchannels frequently used by the AeNB. Firstly, there is no overlap in the spectrum used by the AeNB and the PeNBs. Secondly, the AeNB uses fewer subchannels (less spectrum), since the small cells successfully avoid using a number of the AeNB's most preferred subchannels. This in turn positively reinforces the use of the same subchannels by the AeNB through the Q-learning algorithm.

C. Primary User Quality of Service

Fig. 7 shows contour plots of the spatial distribution of user throughput (UT) across the area outside of the stadium, covered by the PeNBs and the AeNB. They indicate that the area most susceptible to harmful interference is that in the vicinity of the stadium, where the UEs are connected to the AeNB as well as the PeNBs. There is also interference radiating from the ultra-dense stadium small cell network. Fig. 7a shows that the "dynamic ICIC" approach, with a relatively even spectrum occupancy distribution seen in Fig. 6a, performs poorly and results in a significant decrease in UT in the vicinity of the stadium. Such performance degradation of the UEs located outside of the stadium is unacceptable from the viewpoint of secondary spectrum sharing. A significant improvement in the spatial UT distribution is achieved by using the learning based "DIAQ + Q-learning" approach. The performance is further improved by using the novel "DRIAQ + RAQ-AC" approach proposed in this chapter due to its ability to autonomously achieve the significantly more adaptable spectrum partitioning patterns seen in Fig. 6c.

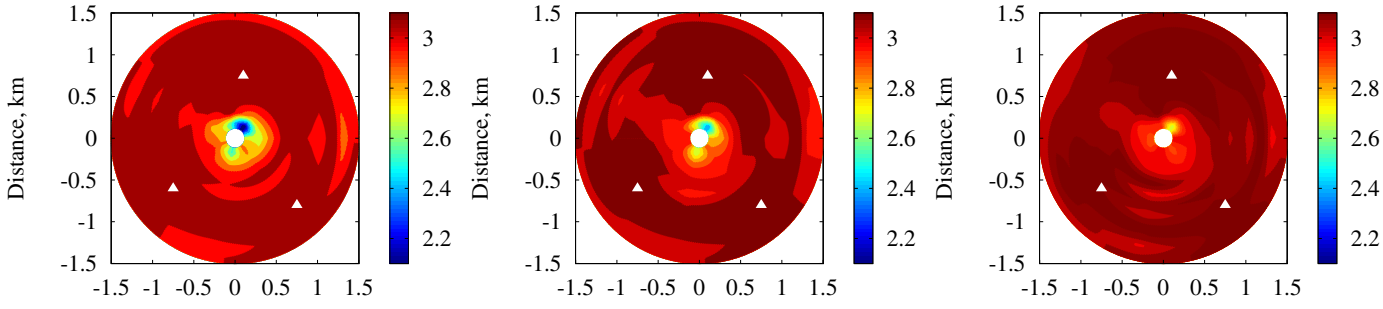
D. Statistical Analysis

The results in Fig. 8 break down the QoS provided to the primary and secondary system users using the three different DSS strategies. Furthermore, they also verify the statistical significance of performance improvements gained by using the HARL based "DRIAQ + RAQ-AC" scheme proposed in Section IV. It shows the results from 50 different simulation setups, i.e. with different random seeds, UE locations and initial traffic, in the form of box plots [36], a compact way of depicting key features of probability distributions.

Fig. 8a shows that the variation in mean UT outside the stadium is negligibly small, when comparing different DSS strategies. The equation for calculating UT for any given UE, as defined in [33], is given below:

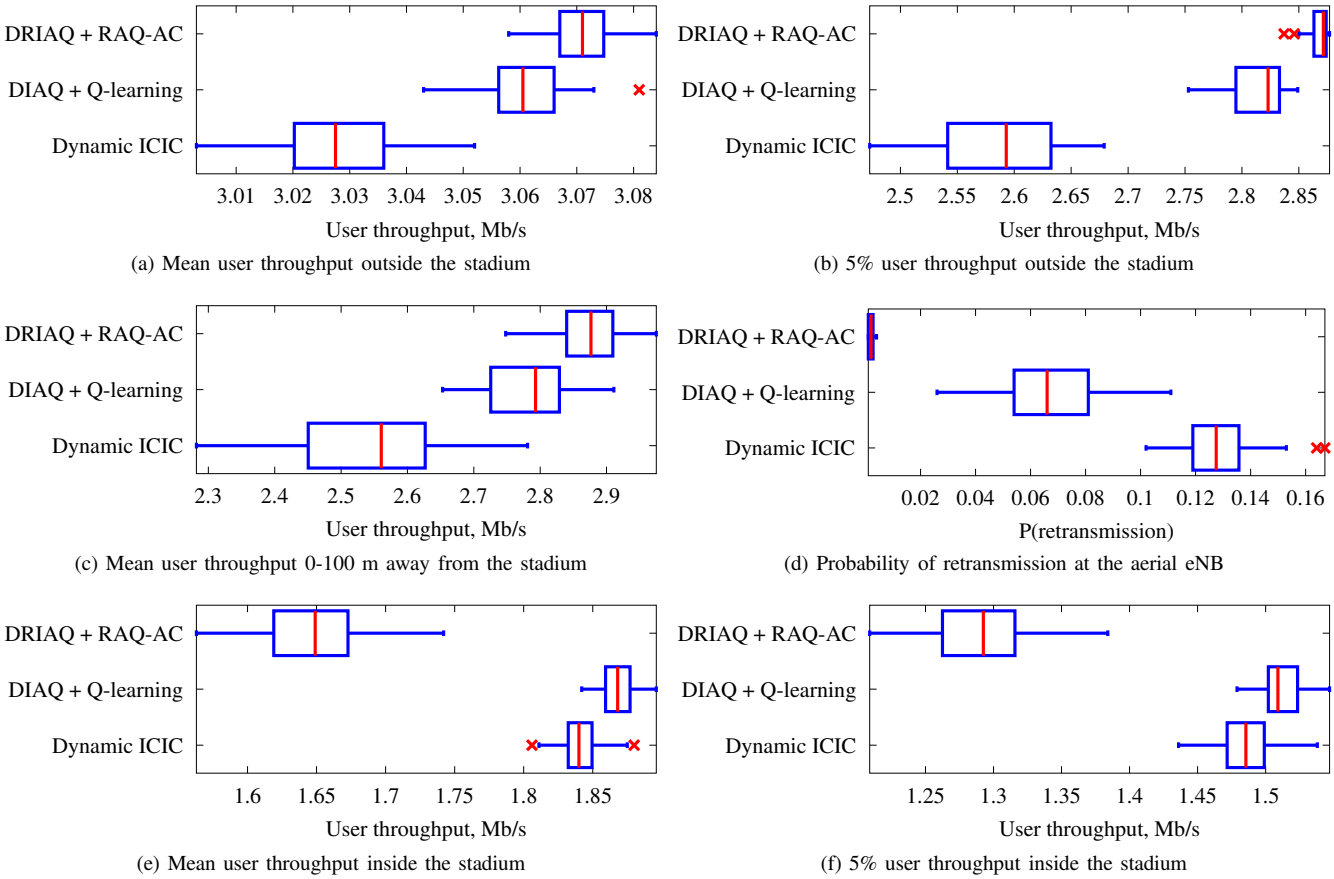
$$UT = \frac{\sum_{f=1}^{F} S_f}{\sum_{f=1}^{F} T_f} \tag{19}$$

where F is the number of files downloaded by the given UE, S_f is the size of the f^{th} file, and T_f is the time it took to download it.



(a) Dynamic ICIC (b) DIAQ + Q-learning (c) DRIAQ + RAQ-AC

Fig. 7. Spatial distribution of user throughput (Mb/s) outside of the stadium (the triangles represent the primary eNB locations)



(a) Mean user throughput outside the stadium

(b) 5% user throughput outside the stadium

(c) Mean user throughput 0-100 m away from the stadium

(d) Probability of retransmission at the aerial eNB

(e) Mean user throughput inside the stadium

(f) 5% user throughput inside the stadium

Fig. 8. Boxplots of the primary and secondary system performance from 50 different simulations

However, the box plots of 5% UT outside the stadium in Fig. 8b reveal a more significant difference in the performance of the simulated DSS schemes. 5% UT for a single simulation is obtained by calculating the 5th percentile of the UT values of 500 users outside the stadium. It is a more important metric than the mean UT, since it represents a minimum QoS guaranteed to 95% of the users, and thus shows how fair the spatial QoS distribution is. Introducing the learning algorithms into the spectrum sharing strategies ("DIAQ + Q-learning") results in an 8.9% increase in median 5% UT outside the stadium compared to "dynamic ICIC", whereas the novel "DRIAQ + RAQ-AC" scheme improves it by 11%. These improvements are statistically significant since there is no overlap between the boxes in the plot. The same improvement pattern is observed in Fig. 8c which shows the mean UT of the users located in the vicinity of the stadium (0-100m from the boundary), the region most vulnerable to the interference between the small cell network, the AeNB and the PeNBs.

Fig. 8d demonstrates the most notable performance improvement achieved by "DRIAQ + RAQ-AC". It almost entirely eliminates the retransmissions, i.e. the blocked and interrupted file transmissions, at the AeNB. It results in a 98% decrease in the probability of retransmission $P(re-tx)$ compared to "dynamic ICIC" and a 97% decrease compared to a significantly better "DIAQ + Q-learning" scheme. $P(re-tx)$ is defined as the ratio between the number of retransmissions and the total number of transmissions. This improvement is achieved due to high controllability provided by the the heuristic functions designed in Section IV. They successfully steer the learning process of the AeNB such that it avoids interfering with the PeNBs, whereas the small cell eNBs are continuously discouraged from occupying the resources preferred by the AeNB, as demonstrated by the spectrum occupancy patterns in Fig. 6c.

Fig. 8e and 8f show that the improvements in QoS, provided by the "DRIAQ + RAQ-AC" scheme to the PeNB and AeNB users, come at the cost of a 10-12% decrease in mean UT and a 13-14% decrease in 5% UT provided to the small cell users, compared with the two baseline schemes. However, this concession made by the stadium small cell network is relatively insignificant and essential in the context of dynamic secondary spectrum sharing. It results in the increased feasibility of secondary LTE spectrum reuse by a temporarily deployed eNB on an aerial platform and an ultra-high capacity density stadium small cell network, that is able to accommodate a vast increase in capacity (1 Gb/s in addition to the primary system's 20 Mb/s offered traffic). Furthermore, the "DRIAQ + RAQ-AC" scheme achieves remarkable reliability of AeNB

communications (due to the lack of retransmissions). For example, this could be highly useful in the temporary event scenario for providing a robust dedicated access network to event organizers both inside and outside the stadium.

E. Temporal Performance

Fig. 9 shows the temporal performance of the two learning based schemes, "DIAQ + Q-learning" and "DRIAQ + RAQ-AC", in terms of the probability of retransmission at the AeNB. All data points were obtained by averaging over 50 different simulations. The time response of "DIAQ + Q-learning" demonstrates that it behaves as a classical RL algorithm, i.e. starts at a relatively poor performance level and gradually improves over time, while the AeNB and the small cell eNBs are learning appropriate spectrum sharing patterns. In contrast, the "DRIAQ + RAQ-AC" time response is a great demonstration of the temporal performance improvements achieved by introducing heuristic acceleration into the learning process. It starts at a superior $P(re-tx)$ level and maintains it throughout the whole simulation.

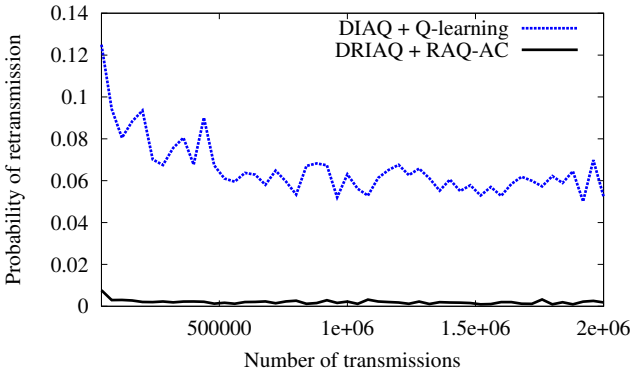


Fig. 9. Probability of retransmission time response at the aerial eNB

VI. Conclusion

The HARL based framework proposed in this paper utilizes a radio environment map (REM) as external information for guiding the learning process of cognitive cellular systems, which are thus able to reuse the LTE spectrum owned by another cellular network. The performance of the DSS and DSA schemes investigated in this paper is assessed using system level simulations of a stadium temporary event scenario. This involves an eNodeB on an aerial platform, a small cell stadium network and a local primary LTE network. Two novel DSS schemes are described in detail - distributed REM and ICIC accelerated Q-learning (DRIAQ) used by the small cell network, and REM accelerated Q-learning with Q-value based admission control (RAQ-AC) used by the aerial eNodeB. These schemes are shown to achieve high controllability of spectrum sharing patterns in a fully autonomous way. They also result in a significant decrease in primary system QoS degradation due to the interference from the secondary cognitive systems, compared to a state-of-the-art RL solution and a purely heuristic typical LTE solution. The spectrum sharing patterns that emerge by using the proposed schemes also result in remarkable reliability of the cognitive aerial eNodeB due to a 97% decrease in the probability of retransmission compared to a classical RL approach.

Furthermore, the novel principle of superimposed heuristic functions proposed in the context of HARL, as well as the general Q-table mask structure of these functions, are not specific to the investigated spectrum sharing scenario, and are generally applicable to a wide range of self-organization problems beyond the wireless communications domain.

Acknowledgements

This work was funded by the ABSOLUTE Project (FP7-ICT-2011-8-318632), which receives funding from the 7th Framework Programme of the European Commission.

References

[1] H. Sun, A. Nallanathan, C.-X. Wang, and Y. Chen, "Wideband spectrum sensing for cognitive radio networks: a survey," *Wireless Communications, IEEE*, vol. 20, pp. 74–81, 2013.

[2] M. Guizani, B. Khalfi, M. Ghorbel, and B. Hamdaoui, "Large-scale cognitive cellular systems: resource management overview," *Communications Magazine, IEEE*, vol. 53, pp. 44–51, 2015.

[3] C. Ghosh, S. Roy, and D. Cavalcanti, "Coexistence challenges for heterogeneous cognitive wireless networks in TV white spaces," *Wireless Communications, IEEE*, vol. 18, pp. 22–31, 2011.

[4] D. Gurney, G. Buchwald, L. Ecklund, S. Kuffner, and J. Grosspietsch, "Geo-location database techniques for incumbent protection in the TV white space," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2008.

[5] M. Matinmikko, H. Okkonen, M. Palola, S. Yrjola, P. Ahokangas, and M. Mustonen, "Spectrum sharing using licensed shared access: the concept and its workflow for LTE-advanced networks," *Wireless Communications, IEEE*, vol. 21, pp. 72–79, 2014.

[6] S. Hamouda, M. Zitoun, and S. Tabbane, "Win-win relationship between macrocell and femtocells for spectrum sharing in LTE-A," *Communications, IET*, vol. 8, pp. 1109–1116, 2014.

[7] G. Alnwaimi, T. Zahir, S. Vahid, and K. Moessner, "Machine Learning based Knowledge Acquisition on Spectrum Usage for LTE Femtocells," in *IEEE Vehicular Technology Conference (VTC-Fall)*, 2013.

[8] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[9] T. Jiang, D. Grace, and P. D. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET*, vol. 5, pp. 1309–1317, 2011.

[10] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and H. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *Wireless Communications, IEEE Transactions on*, vol. 12, pp. 3202–3212, 2013.

[11] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems," *Mobile Computing, IEEE Transactions on*, vol. PP, pp. 1–11, 2015.

[12] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *Mobile Computing, IEEE Transactions on*, vol. 12, pp. 2155–2166, 2013.

[13] C. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge, England, 1989.

[14] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *Vehicular Technology, IEEE Transactions on*, vol. 48, pp. 1676–1687, 1999.

[15] R. Valcarce, T. Rasheed, K. Gomez, S. Kandeepan, L. Reynaud, R. Hermenier, A. Munari, M. Mohorcic, M. Smolnikar, and I. Bucaille, "Airborne base stations for emergency and temporary events," in *International Conference on Personal Satellite Services*, 2013.

[16] R. Bianchi, M. Martins, C. Ribeiro, and A. Costa, "Heuristically-accelerated multiagent reinforcement learning," *Cybernetics, IEEE Transactions on*, vol. 44, pp. 252–265, 2014.

[17] R. Bianchi and R. Lopez de Mantaras, "Case-based multiagent reinforcement learning: Cases as heuristics for selection of actions," in *European Conference on Artificial Intelligence (ECAI 2010)*, 2010.

[18] J. Lunden, S. Kulkarni, V. Koivunen, and H. Poor, "Multiagent reinforcement learning based spectrum sensing policies for cognitive radio networks," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, pp. 858–868, 2013.

[19] S. Chen, R. Vuyyuru, O. Altintas, and A. M. Wyglinski, "On optimizing vehicular dynamic spectrum access networks: Automation and learning in mobile wireless environments," in *Vehicular Networking Conference (VNC), 2011 IEEE*, 2011.

[20] L. Reynaud, et al., "FP7-ICT-2011-8-318632-ABSOLUTE/D2.1 Use cases definition and scenarios description," 2014.

[21] N. Morozs, D. Grace, and T. Clarke, "Distributed Q-learning based dynamic spectrum access in high capacity density cognitive cellular systems using secondary LTE spectrum sharing," in *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2014.

[22] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998.

[23] N. Morozs, T. Clarke, and D. Grace, "A novel adaptive call admission control scheme for distributed reinforcement learning based dynamic spectrum access in cellular networks," in *International Symposium on Wireless Communication Systems (ISWCS)*, 2013.

[24] N. Morozs, T. Clarke, D. Grace, and Q. Zhao, "Distributed Q-learning based dynamic spectrum management in cognitive cellular systems: Choosing the right learning rate," in *IEEE International Symposium on Computers and Communications (ISCC)*, 2014.

[25] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, pp. 215–250, 2002.

[26] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 11.5.0 Release 11) ," Dec. 2013.

[27] S. Sesia, M. Baker, and I. Toufik, *LTE-The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, 2011.

[28] R. McLean, M. Silvius, K. Hopkinson, B. Flatley, E. Hennessey, C. Medve, J. Thompson, M. Tolson, and C. Dalton, "An architecture for coexistence with multiple users in frequency hopping cognitive radio networks," *Selected Areas in Communications, IEEE Journal on*, vol. 32, pp. 563–571, 2014.

[29] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, pp. 276–280, 1986.

[30] G. Ross, N. Adams, D. Tasoulis, and D. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern Recognition Letters*, vol. 33, pp. 191 – 198, 2012.

[31] A. Ghasemi and E. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," *Communications Magazine, IEEE*, vol. 46, pp. 32–39, 2008.

[32] P. Kyösti, et al., "IST-4-027756 WINNER II Deliverable D1.1.2: WINNER II channel models," 2008.

[33] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA physical layer aspects (3GPP TR 36.814 version 9.0.0 Release 9)," Dec. 2010.

[34] ——, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification (3GPP TS 36.321 version 11.4.0 Release 11)," Jan. 2014.

[35] ——, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios (3GPP TR 36.952 version 11.0.0 Release 11)," Dec. 2012.

[36] R. McGill, J. Tukey, and W. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, pp. 12–16, 1978.

Nils Morozs (S'13) received his MEng degree in Electronic Engineering from the University of York in 2012. He is currently a PhD candidate at the University of York and a researcher in Wi-Fi & wireless convergence at BT. His PhD research was part of the EU FP7 ABSOLUTE project concerned with developing novel LTE-compliant cognitive mechanisms for dynamic radio resource and topology management in disaster relief and temporary event networks. His research interests lie in the area of intelligent wireless networks for 5G and beyond.

Tim Clarke received the B.A. degree in biology from the University of York in 1975. He joined the Royal Air Force as an Air Traffic Control Officer before becoming an Education Officer. He underwent advanced training at the Royal Military College of Science, Shrivenham, where he received the M.Sc. degree in guided weapons systems engineering. He is Senior Lecturer in Control Engineering and is Head of the Control Systems Laboratory, Intelligent Systems Group, Department of Electronics, University of York. His research interests are in the areas of biologically inspired engineering and control systems. Mr. Clarke is a member of the IET and serves on IFAC Technical Committees 5.4 (Large Scale Complex Systems), 7.3 (Aerospace), and 7.5 (Intelligent Autonomous Vehicles).

David Grace (S'95-A'99-M'00-SM'13) received his PhD from University of York in 1999, with the subject of his thesis being Distributed Dynamic Channel Assignment for the Wireless Environment. Since 1994 he has been a member of the Department of Electronics at York, where he is now Professor (Research) and Head of Communications and Signal Processing Research Group. He is also a Co-Director of the York - Zhejiang Lab on Cognitive Radio and Green Communications, and a Guest Professor at Zhejiang University.Current research interests include aerial platform based communications, cognitive green radio, particularly applying distributed artificial intelligence to resource and topology management to improve overall energy efficiency; 5G system architectures; dynamic spectrum access and interference management. He is currently a lead investigator on H2020 MCSA 5G-AURA. He was a one of the lead investigators on FP7 ABSOLUTE and focussed on extending LTE-A for emergency/temporary events through application of cognitive techniques. He was technical lead on the 14-partner FP6 CAPANINA project that dealt with broadband communications from high altitude platforms. He is an author of over 220 papers, and author/editor of 2 books. He is the former chair of IEEE Technical Committee on Cognitive Networks for the period 2013/4. He is a founding member of the IEEE Technical Committee on Green Communications and Computing. In 2000, he jointly founded SkyLARC Technologies Ltd, and was one of its directors. He is currently a Non-Executive Director of a technology start-up company.